A CONTEXT-AWARE INTELLIGENT SYSTEM TO ASSIST USER PROFILEFILTERING USING AI AND DEEP LEARNING

Xinrui Que[1] and Yao Pan[2]

[1]Crean Lutheran High School, 12500 Sand Canyon Ave,
Irvine, CA 92618, USA
[2]Department of Computer Science, Vanderbilt University, USA

ABSTRACT

Community based websites such as social networks and online forums usually require users to register by providing profile information and avatars. It is important to ensure these user uploaded information comply with the website policy. This includes the information being personal, related and clear, as well as not containing unhealthy/disturbing content. A review or censorship system is usually deployed to review new user registration. Nowadays, many platforms still use manual review or rely on 3rd party APIs. However, manual review is time-consuming and costly. While 3rd party services are not tailored to the specific business needs thus do not provide enough accuracy.

In this paper, we developed an automatically new user registration review system with deep learning. We apply the state-of-art techniques such as CNN and BERT for an end-to-end evaluation system for multi-modal content. We tested our system in E-pal, a freelancing platform for gaming companionship and conducted a qualitative evaluationof the approach. The results show that our system can evaluate the quality of avatars, voice descriptions, and text profiles with high accuracy. The system can significantly reduce the effort of manual review and also provides input for the recommendation ranking.

KEYWORDS

Deep learning, Image classification, BERT, CNN.

1. INTRODUCTION

Community based websites such as social networks and online forums usually require users to register by providing profile information and avatars. It is important to ensure these user uploaded information comply with the website policy. This includes the information being personal, related and clear, as well as not containing unhealthy/disturbing content. A review or censorship system is usually deployed to review new user registration. Nowadays, many platforms still use manual review or rely on 3rd party APIs. However, manual review is time-consuming and costly. While 3rd party services are not tailored to the specific business needs thus do not provide enough accuracy.

In this paper, we focus on a specific user case: E-pal.gg [1], which is a freelancing platform for gaming companionship. But the framework can be applied to other platforms/websites as well. The E-pal community is composed of E-pal and gamer. E-pal gets commissions by playing games with other gamer or teaching others to play games. The platform, as an intermediary,

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 65-75, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120105

connects E-pals and gamer. When an E-pal registers an account on the platform, they are asked to use their photos as avatars, speak a language describing themselves, and write a self-descriptive text. We need to ensure that the avatar is personal and clear. There is no racial discrimination in language and writing, no sexual suggestion, and no unhealthy content.

Nowadays, many platforms still use manual review, where some contractors are hired to perform evaluation tasks and decide whether the image/text comply with policy. However, there are several limitations with manual review: 1) manual review is slow. Manual review takes significant time and adds operational cost. The member base of popular platforms can easily get to millions or even billions of users. For many start-ups, they could also face a user growth explosion where thousands of new users register every day. It could require multiple people. 2). manual review is not accurate. When a person repeats a simple action, they will get tired and make mistakes. Auditing is a process that is easy to make people tired and make mistakes. Denying qualified registration impacts the platform's user growth. Passing unqualified registration could damage company reputation and cause a serious public relation crisis. 3). Manual review is very subjective. It is easy to have personal emotions in the review and fail to maintain a consistent standard.

Besides manual review, There are companies/websites providing 3rd party image/video/audio evaluation services. For example, Amazon has a Rekognition API. These services are general purpose and train on millions of images across many categories. They work well for general image classification but they are not tailored to the specific business needs thus do not provide enough accuracy.

Open Problem: Review new user registration information in multi-modal (image/audio) accurately and scalable to enable a fast and healthy platform development.

Solution: Automatically new user registration information review with deep learning.

In this paper, we developed an automatic user registration review system with deep learning. Deep learning [2] has received great success in recent years in the domain of computer vision [3] and natural language processing [4]. AI has shown to outperform humans in various tasks such as image classification [5], question answering and reading comprehension [6]. The advancement of deep learning comes from both algorithm improvement such as more advanced network structure, as well as hardware advancement, where the scale and parameters of the model can grow very large.

One particular important advancement for deep learning is transfer learning. Where a model is trained on a large number of labeled data first. The model will learn some general patterns and is then fine-tuned on a specific task.

For our new registration review tasks, we are dealing with multi-modal data from image, text to audio. We apply the state-of-art technique such as MobileNet and BERT to learn an embedding first and then use supervised learning to fine-tune it on our specific task.

Compared to manual review, AI based review systems are more scalable since the system can be easily extended by leveraging cloud computing service. The AI system is also more subjective since it has a unified standard.

In order to evaluate our system, we conducted quantitative experiments to evaluate the effectiveness and accuracy of our proposed solution. First, we experiment with the profile image evaluation where the task is 1) Detect if the image contains any sexual/unhealthful content 2) detect if the image contains a face and the quality of the image in terms of lighting, clearness, Aesthetic, etc. Our system will generate a score from 0 to 100 as well as a reason. The score was compared with a manually reviewed score to calculate the mismatch.

Second, we experiment with audio intro evaluation. The task is 1) Evaluate the audio quality in terms of noise level, speech speed, clearness, appeal/aesthetic value. 2) Detect if the speech content is relevant to the context and does not contain any racial/offensive language. In the end, We also picked some samples for qualitative analysis and demonstrated the effectiveness of the system.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges thatwe met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

Currently, we rely on humans to manually review users' uploaded profiles to determine if they are eligible. However, there are several challenges with the existing approach.

2.1. Review needs to be scalable and fast to keep up with the growth of the platform

Manual review takes significant time and adds operational cost. The member base of popular platforms can easily get to millions or even billions of users. For many start-ups, they could also face a user growth explosion where thousands of new users register every day. It could require multiple people.

Also it's hard to predict the user growth. If we hire more people but not enough registration, we add unnecessary operational cost, Or if the user growth outpaces the human review, the review will be delayed and lead to poor member experience.

2.2. Review needs to be accurate

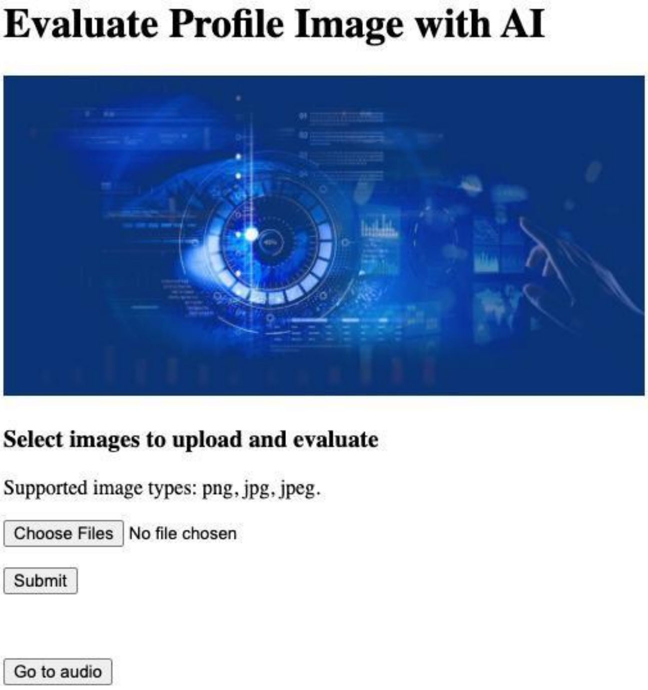
We need to minimize the probability of denying qualified registration as well as passing unqualified registration. Manual review is not accurate as humans make mistakes. Review is a repetitive and tedious process where it's easy for a person to get tired and make mistakes. Denying qualified registration impacts the platform's user growth. Passing unqualified registration could damage company reputation and cause a serious public relation crisis.

2.3. Review needs to be subjective and maintain a consistent standard

Manual review is very subjective and the criteria differs from person to person. On one hand, this inconsistency creates inaccuracy, on the other hand, this could also confuse users if their registration is rejected but similar or even worse registration has passed.

3. SOLUTION

The overall system consists of a front end which is a website built with HTML/JSS, and a back-end which is powered by Flask and Tensorflow. Users will be able to upload images or audio file and the evaluation results will be given for each uploaded item.



Evaluate Profile Image with AI

Select images to upload and evaluate

Supported image types: png, jpg, jpeg.

Figure 1. Overview of the system

The system consists of two sub-systems: profile image scoring system and audio introduction scoring system.

Profile image scoring system. The system has the following components:

1. Image pre-process. In this step, images of various formats and dimensions will need to be converted to a unified format to facilitate following processing.
2. Nudity Content detection. To ensure a healthy platform environment, images with explicit sexual content are not allowed. We utilize an existing library NudeNet (https://github.com/notAI-tech/NudeNet) to help us with nudity censoring.
3. Face detection. Required by the platform, profile images should contain the selfie of the user himself. Sometimes users will upload arbitrary images of landscape, object or cartoon. We perform face detection to distinguish those that have human faces versus those that don't.
4. Image quality evaluation. This subsystem is to evaluate the quality of the profile image. The quality here refers to lighting, clearness, Aesthetic, etc. Here we use a pretrained model efficientnetv2-s to get the embedding of the image. Then we feed the embedding into a deep neural network to train a regression model.

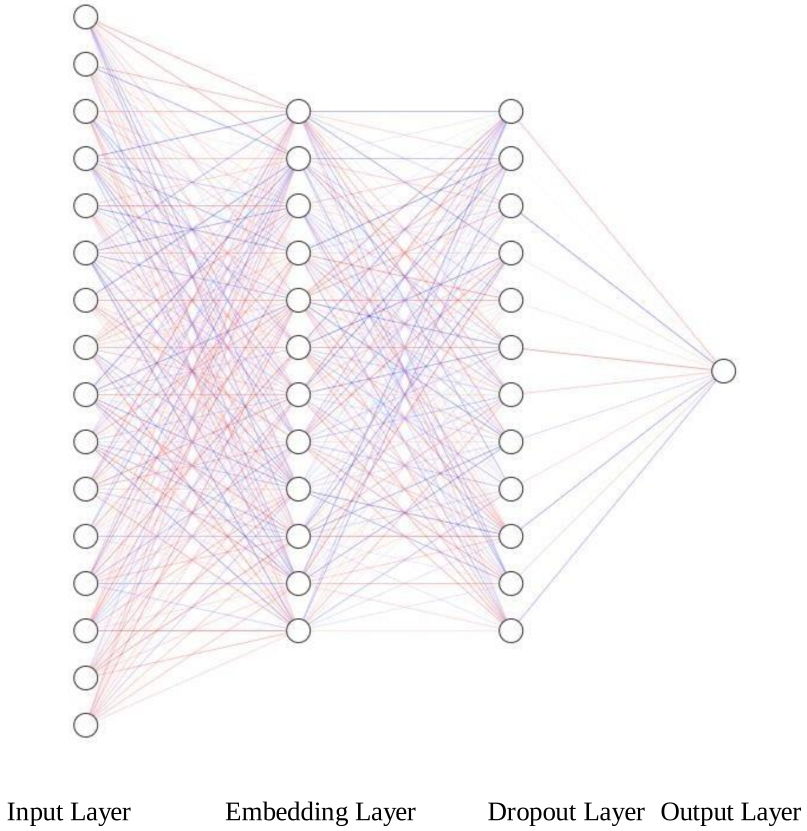


Figure 2. Image classification

Deep learning has received great success in recent years in various tasks such as image classification, object detection, etc. One important advancement is transfer learning. Where a model is trained on a large number of labeled data first. This is very helpful as image models can easily have millions of parameters and require a lot of computing power to rain. The model will learn some general patterns and is then fine-tuned on a specific task. In our case, efficientNet is trained on ImageNet for general image classification. The embedding is then used to fine-tune for profile image scoring.

We rely on efficient v2-s (EfficientNetV2: Smaller Models and Faster Training) to learn an embedding of image. EfficientNet is a new family of CNN with faster speed and better parameter efficiency. It was able to achieve comparable performance compared to some large scale models while remaining small in size and fast serving speed.

Our neural network has 4 layers.

1. The first layer is input layer: it has 384*384*3 dimensions. Where 384*384 is the image sizeand 3 is the color channel.
2. The second layer is the embedding layer. efficient v2-s will return an embedding of dimension1280.
3. The third layer is the dropout layer. It is used for preventing overfitting. A certain percentageof connection is randomly dropped during training.
4. The fourth layer is the output layer. It has a dimension of 1. It outputs a score of 0-100.

Audio introduction scoring system. The system has the following components:

1. Audio preprocess. In this step, images of various format dimensions will need to be converted to a unified format to facilitate following processing. We use ffmpeg which is a widely used audio conversion tool.
2. Speech recognition. We rely on Google cloud Speech recognition API.
3. Text evaluation. We trained a Bert based model.
4. Audio evaluation. This subsystem is to evaluate the quality of the intro audio. The quality here refers to noise level, speech speed, clearness, appeal/aesthetic value.

For the text-based evaluation, we build a neural network based on BERT [7] embedding. BERT (Bidirectional Encoder Representations from Transformers) is a transformer based model which has been widely successful on a variety of NLP tasks such as text classification, sentiment analysis, questions answering, machine translation, etc [8].

The text from the speech recognition API is fed into the input of the neural network. Then there is a pre processing layer which computes the input_words_id, input_mask and input_type_ids. The three inputs are passed to the BERT encoder, and output a 1024 length vector. We then include a 512 length dense layer with relu activation function. Anally an output layer of length 2. The network diagram is shown in Fig 3.

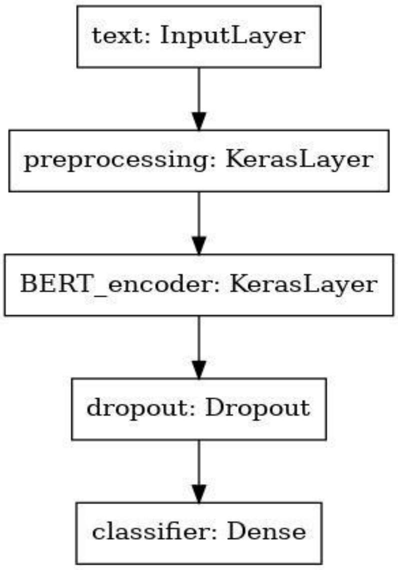


Figure 3. Network diagram

Although the BERT model can capture the content of the audio and determine whether it is relevant, it cannot capture other factors of audio such as whether the voice is clear, speech speed is appropriate or the voice is appealing. So we build a second neural network to evaluate the voice features of the audio. We use YAMNET [9] to generate embedding of the audio and usethe embedding to train a classifier to predict whether the audio is high quality or low quality.

YAMNET is a deep network that predicts audio from 521 classes. It returns a score vector indicating the probability for each of the 521 classes. It also returns an embedding of shape (N, 1024) where N is the number of 0.96 second frames. We conducted an average-pooling to get a vector of 1024.

4. EXPERIMENT

In this section, we described experiments to evaluate the effectiveness and accuracy of our proposed solution.

We collected 2137 real user profile avatar and audio from E-pal. Both image and audio are manually reviewed and given a score from 0 to 100. The score distribution is image and audio are given in Fig 4 and 5. We use 80% of the data as training and the remaining 20% as evaluation.

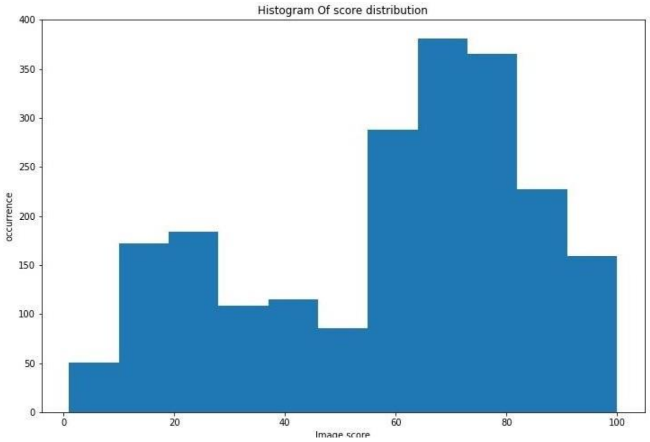


Figure 4. Histogram of image score distribution

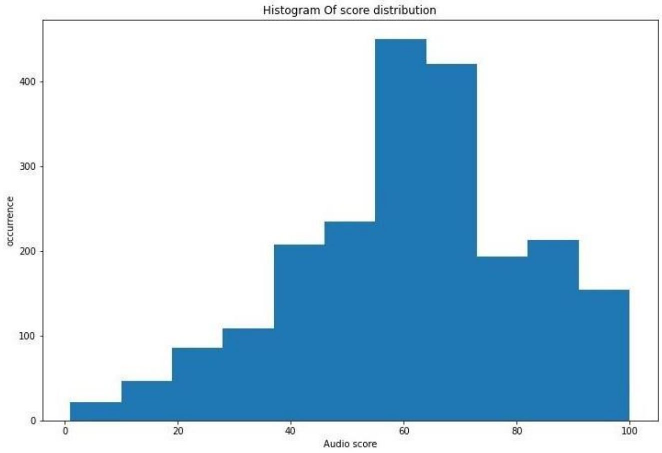


Figure 5. Histogram of audio score distribution

1. Evaluate Profile Image Scoring System.

First, we experiment with the profile image evaluation where the tasks is 1) Detect if the image contain any sexual/unhealthy content 2) detect if the image contain face and the quality of the image in terms of lighting, clearness, Aesthetic, etc. Our system will generate a score from 0 to 100 as well as a reason. The score was compared with a manually reviewed score to calculate the mismatch.

We use two metrics to evaluate our results. 1. mean absolute error (MAE). MAE is defined as

100ni=1n|Ai-Fi|. At is the actual score and Fi is the predicted score. This is to get a sense of how close the predicted score is compared to actual score. 2. Accuracy. Here we treat the problem as a classification problem where we only have 3 classes: High quality (score above 70), medium quality (40-70) and low quality (score below 40). The metric is used to give a rough estimation of the evaluation quality.

The hyper-parameters are set as follows:

learning rate = 0.005, momentum =0.9, L2 regularization = 0.0001, batch size =16, dropout rate =0.2.

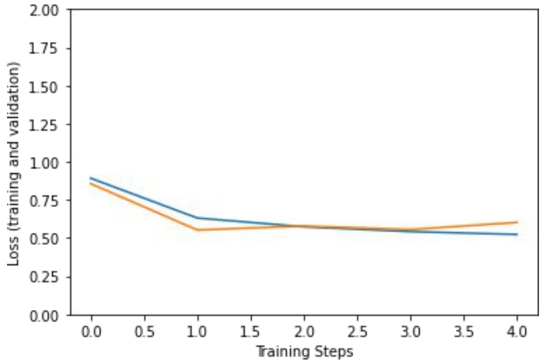


Figure 6. Loss vs. Training steps

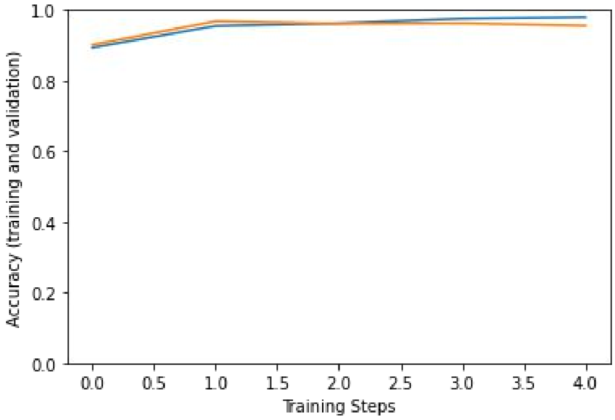


Figure 7. Accuracy vs. Training steps

Table 1. Performance of image evaluation on different embedding

Image embedding	MAE	Accuracy
Efficientnet v2-s	12.4	91.2%
Efficientnet v2-m	12.2	91.5%

In table 1, we compare the performance of two embedding (efficient netv2-m has a larger dimension than efficient netv2-s). We can see that both can accurately evaluate the image. The absolute score difference is around 12 and over 90% of the samples are correctly classified into the high/medium/low classes. Having a larger embedding model slightly improves the performance, but also at the cost of longer training and scoring execution time.

We also look at the images where the predicted score and labeled score differs the most. We found this most happened for certain cartoon images where the cartoon character is very human. The algorithm tends to give them higher scores while human reviewers give lower scoresbecause they are not the photo of users. One of our future work would be to improve the algorithm to better distinguish cartoon characters.

We also tested the performance of scoring with the train model. We deployed the tensor-flow model on an Amazon ec2 t3.large instance. The average scoring time is 0.83s, which means that it can score over 100000 images per day. And the cost is only 0.0832*24=$2 per day, which is much lower than human reviewers.

2. Evaluate Audio Intro Scoring System.

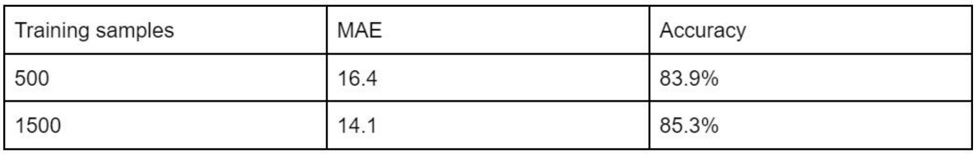
For the audio intro evaluation, the task is 1) Evaluate the audio quality in terms of noise level, speech speed, clearness, appeal/aesthetic value. 2) Detect if the speech content is relevant to the context and does not contain any racial/offensive language.

For the BERT model, the hyperparameters are set as follows:

learning rate = 2e-5, batch size = 32, max sequence length = 64, epsilon = 1E-8.For audio neural network, the hyperparameters are set as follows:

learning rate = 0.005, momentum =0.9, L2 regularization = 0.0001, batch size =16, dropout rate =0.2.

Table 2. Performance of audio evaluation on different training sample size



Training samples	MAE	Accuracy
500	16.4	83.9%
1500	14.1	85.3%

We can see that the proposed solution can also accurately evaluate the audio. The absolute score difference is around 14 and over 85% of the samples are correctly classified into the high/medium/low classes. The performance improves as we increase the training sample size (from 500 to 1500). If more training data is available, the accuracy could be further improved.

5. RELATED WORK

Face detection is a topic that has been widely discussed in computer vision over the past few decades. Viola. et al [10] propose a detection framework based on Haar features and Adaboost classifier. In recent years, deep learning has achieved great success in many computer vision tasks. Deep convolutional neural net based methods [11] [12] have outperformed traditional machine learning methods in face detection in both the accuracy and ease of use.

Image aesthetic assessment [13] aims to computationally distinguish high-quality image from low-quality image based on photographic rules. Different approaches have been developed, some based on hand-crafted features [14] and some based on deep features [15].

Speech or Audio quality assessment has been studied by several researchers [16]. Some rely on signal-to-noise ratio measures. Some rely on spectral distance measures. However, these low level features cannot capture the semantic meaning of the speech. Text classification is the tasks of classifying text into multiple classes based on the semantic meaning. The applications range from email spam classification [17], sentiment classification, There are companies/websites providing 3rd party image/video/audio evaluation services. For example, Amazon has Rekognition API [18]. These services are general purpose and train on millions of images across many categories. They work well for general image classification but are not tailored to specific business needs.

6. CONCLUSIONS

In this paper, we proposed an automatically new user registration review system with deep learning. We apply the state-of-art techniques such as CNN and BERT for an end-to-end evaluation system for multi-modal content. The system can be used in various Community based websites such as social networks and online forums. We conducted an experiment using real world profile data from E-pal, a freelancing platform for gaming companionship. The results indicate deep learning can accurately classify low quality profile vs. high quality profile.

There are still some limitations with the current approach. The algorithm generally performs well at identifying low quality or high quality input, but accuracy can become lower for border line cases. Also, there are some corner cases not well covered by the current algorithm. For example, cartoon characters are sometimes still detected as human faces.

For future work, we plan to continue to improve the algorithm accuracy by cleaning andobtaining more training data, experiment with more complex models and improve the explainability of the scoring results.

REFERENCES

[1] E-pal. https://www.epal.gg/

[2] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[3] Voulodimos, Athanasios, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. "Deep learning for computer vision: A brief review." Computational intelligence and neuroscience 2018 (2018).

[4] Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. "Recent trends in deep learning based natural language processing." ieee Computational intelligenCe magazine 13, no. 3 (2018): 55-75.

[5] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human- level performance on imagenet classification." In Proceedings of the IEEE international conference on computer vision, pp. 1026-1034. 2015.

[6] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683 (2019).

[7] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[8] Young, Tom, et al. "Recent trends in deep learning based natural language processing." ieee Computational intelligence magazine 13.3 (2018): 55-75.

[9] https://tfhub.dev/google/yamnet/1

[10] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001,vol. 1, pp. I-I. Ieee, 2001.

[11] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." In 2017 International Conference on Engineering and Technology (ICET), pp. 1-6. Ieee, 2017.

[12] Sun, Xudong, Pengcheng Wu, and Steven CH Hoi. "Face detection using deep learning: An improved

fasterRCNN approach." Neurocomputing 299 (2018): 42-50.

[13] Deng, Yubin, Chen Change Loy, and Xiaoou Tang. "Image aesthetic assessment: An experimental survey."IEEE Signal Processing Magazine 34, no. 4 (2017): 80-106.

[14] Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Z. Wang. "Studying aesthetics in photographic images usinga computational approach." In European conference on computer vision, pp. 288-301. Springer, Berlin, Heidelberg, 2006.

[15] Peng, Kuan-Chuan, and Tsuhan Chen. "Toward correlating and solving abstract tasks using convolutional neural networks." In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1-9. IEEE, 2016.

[16] Loizou, Philipos C. "Speech quality assessment." In Multimedia analysis, processing and communications, pp. 623-654. Springer, Berlin, Heidelberg, 2011.

[17] Abdulhamid, Shafi'I. Muhammad, Maryam Shuaib, Oluwafemi Osho, Idris Ismaila, and John K. Alhassan. "Comparative Analysis of Classification Algorithms for Email Spam Detection." International Journal of Computer Network & Information Security 10, no. 1 (2018).

[18] Amazon Rekognition. https://aws.amazon.com/rekognition/